The impact of educational technology:

a radical reappraisal of research methods

P. David Mitchell

Graduate Programme in Educational Technology, Concordia University, Montreal, Canada

How can we decide whether some new tool or approach is valuable? Do published results of empirical research help? This paper challenges strongly entrenched beliefs and practices in educational research and evaluation. It urges practitioners and researchers to question both results and underlying paradigms. Much published research about education and the impact of technology is pseudo-scientific; it draws unwarranted conclusions based on conceptual blunders, inadequate design, so-called measuring instruments that do not measure, and/or use of inappropriate statistical tests. An unacceptably high portion of empirical papers makes at least two of these errors, thus invalidating the reported conclusions.

Introduction

The practical problem which motivates this paper is that of deciding – on the basis of published research – whether to adopt some new device, procedure or paradigm thought likely to improve education. What models, methods or media are likely to be most useful? From the invention of the printing press to multimedia software, educators have adopted unproven aids and fads. Researchers usually claim each new device or procedure to be at least as effective as its predecessor. How valid is all this research? How to decide? A typical view is: *Design an experiment to observe the effects of your treatment. Any book on research design and statistics will show you how.* But will it? What essential aspects of educational measurement and research must we consider?

Measurement or sorcery?

Suppose we wish to conduct research on media-based learning as a function of different learning styles. Let us assume that we decided operationally to define relevant styles with a commonly used questionnaire (for a more complete discussion of learning styles and problems of identifying them, see Mitchell, 1994). A typical questionnaire asks a series of

questions that one answers on a scale of possible responses ranging from, for example, *strongly agree* to *strongly disagree*. But let us examine a measurement issue first.

Measurement: neglected rules

From mathematics, the theory of numbers and the theory of measurement provide the foundation upon which educational measurement and statistics must rest if the latter are to be more than superficial and deceptive. The axiom of identity requires that: each question be equivalent to each of the others; two people with the same score must have comparable abilities; and equal differences between scores be equivalent. The result, if the assumption of equivalence is not violated, is similar to a thermometer; a one-degree difference is the same unit regardless of the starting temperature. Such equal interval scales (see Stevens, 1946) are common in science but not in education.

Instruments presumed to measure some variable like an attitude, opinion or even knowledge, seldom have equivalent questions. Moreover such technical refinements as reliability, validity or internal consistency fail to satisfy this axiom of identity. There is no guarantee that identical scores represent students with identical answers; indeed, it is very unlikely. Yet researchers usually treat questionnaires dealing with linguistic concepts (for example, comprehension or learning style) as if they were sharply defined interval scales. Most scales used in educational research actually are ordinal scales and therefore do not meet the mathematical preconditions for the statistical manipulations commonly used (Liebetrau, 1983).

Consider the questions on commonly used 'instruments' purported to measure learning styles (there are over 100, but see Entwistle, 1981). Typically in such scales, response categories are ranked in order of importance to the researcher. Ranking may begin with *definitely disagree* assigned a rank of 1, and so on to 5. The test creator usually considers this rank to be a 'score' for each question so that he or she can perform statistical analyses on the numbers. Another conceptual blunder is to add the so-called scores for several questions to get a 'total' for that subscale which carries a label supposedly denoting a variable (for example, a particular learning style). This occurs despite the items' appearing to violate the axiom of identity; thus they cannot be added even if each scale were interval.

With a wave of a magic wand (accompanied by the incantation, 'let us assume . . .') it seems that we can represent a statement ('I disagree that . . .') by a number which is not simply a symbol or identifier of a position in a sequence but a quantity. But can we? Is it justified? Mathematically, the difference between 4 and 3 is equivalent to 2 minus 1 or 3 minus 2, but is it correct to say that the difference between my saying that 'I definitely agree . . .' and 'I agree with reservations . . .' is the same as between 'impossible to give a definite answer' and 'I disagree with reservations . . .'? Logically, all we can assume with a ranking of categories is the sequence. Statistics deals with numbers, not what they represent. If numbers, as collected and assigned, violate epistemological or mathematical requirements, the analysis will produce mathematically correct results. But what do they mean?

Measurement or intellectual pollution?

Many published 'measurement instruments' were generated by factor analysis. Surely this procedure justifies the scale and its scoring? Space limitations permit no discussion here,

but this argument should be interpreted in the light of Patrick Meredith's pithy comments about Spearman's contribution to the topic:

> What is disturbing is that Spearman's 'factorial' concept, whose epistemological basis is riddled with fallacies, not only took off but came to dominate the psychological and educational skies [. . .] Instructional Science has a decontamination job on its hands, to disperse the intellectual pollution created by a whole profession reared on a contempt for real information and a superstitious worship of false quantification (Meredith, 1972, p. 16).

Is it possible that educational researchers have a 'contempt for real information and a superstitious worship of false quantification'?

Information, numbers and statistics

In contrast to Stevens (1946) and his followers, I assert that measurement is not just the assignment of numbers to things according to specified operations. The purpose of measurement is to reduce the variety of some part of reality which we observe, whether directly or through some information-gathering activity, to yield summarizing information that is accurate, precise and general. Usually our intention is to answer a question or to support a decision.

Pseudo-measurement

What too frequently happens is that the 'score' produced by the 'scoring key' (by illegitimately summing ranks of ordinal measures) is treated as if it were quantitative information about that variable for each person. Textbooks and professors often claim that it is all right to treat Ordinal Scales as if they were Interval Scales because their test of significance is so robust that it is unlikely to lead to improper conclusions. How credible is this? Note that 'robust' is contextual, not fixed, contrary to a common myth. And any violation of a test's prerequisites alters its probabilities of Type I and II errors. Moreover, the powers of some parametric tests have been shown to diminish to zero under violations of mathematical assumptions of the test (Bradley, 1982).

If we play games with epistemological underpinnings and mathematical prerequisites, the consequences are unknown, and our analysis could be meaningless. Lakatos, a philosopher of science, dismissed our typical use of statistical techniques to produce 'phoney corroborations and thereby a semblance of "scientific progress" where, in fact, there is nothing but an increase in pseudo-intellectual garbage' (Lakatos, 1978, p. 88).

Insignificance of statistical significance

Consider this quotation from a typical textbook:

> Tests of statistical significance are used to help researchers to draw conclusions about the validity of a knowledge claim. [. . .] If the null hypothesis is rejected, we conclude that the knowledge claim (i.e. the research hypothesis) is true. If the null hypothesis is accepted, we conclude that the knowledge claim is false. (Meehl, 1978, p. 622).

We usually are told to reject the null hypothesis if the difference is 'significant' (i.e. $p<0.05$). But consider Meehl's summarizing statement: '[. . .] if you have enough cases and your measures are not totally reliable, the null hypothesis will always be falsified, regardless of

the truth of the substantive theory'. Despite its prevalence, null-hypothesis significance testing has been criticized for half a century. Lykken's conclusion can guide us:

> Finding of statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published. The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on. (Lykken, 1968, p. 159)

Inappropriate use of parametric statistics

The majority of reported studies reflect the parametric statistical techniques taught in educational research courses for decades. These 'assume' (i.e. have as a mathematical pre-requisite) a normal distribution of the data in a population, even though educational researchers seldom know much about the population beyond the small sample taken in their experiment. (Paradoxically, the more experimental control and the greater the treatment effect, the less approximately normal the distribution will be, and the test's power is attenuated.)

In parametric statistical analysis, what we find is an answer to this question: 'IF the null hypothesis (of no difference between groups) is true, AND if both normality and homogeneity of variance are true, what is the probability that we would find a difference at least as large as that which we observed, if we had randomly drawn two samples from that population?' This is easy to compute, but it is only correct if the prerequisite assumptions are true. Alas, a major problem exists: we do not know if the null hypothesis is true, if the data is normally distributed, if homogeneity of variance is true, or even if the data is random (though our design may reassure as about randomness). So what happens? The researcher can only act as if this were the case by assuming it to be so. Much of the time, it is not.

Simple observation of published means and standard deviations reveals many which clearly cannot be normal distributions (for example, relatively large standard deviations, often nearly as large as, and sometimes several times larger than, the mean; non-symmetrical shape). Lohr *et al* (1995) reported several differences in their 'comprehensive evaluation' of a hypertext model for teaching, but their conclusions came from Analyses of Variance, some of which were 'significant'. Although Meehl (1978) shows that one should not even make such claims with multiple tests, let us examine the comparisons. The authors offer 21 means and standard deviations. Given that a normal distribution is symmetrical and extends three standard deviations above and below the mean, it is noteworthy that only one of the 21 groups could be approximately normal. Eight involved standard deviations as large as or larger than the mean. Clearly these do not describe normally distributed data, and a parametric test should not be used. Unlike those authors, we can conclude nothing except that the analysis was inappropriate. Over half the empirical papers in recent issues of several journals make the same mistake.

As Krauth showed:

> Examination of real data reveals that the assumption of a normal distribution is not justified in the majority of cases. Empirical distributions are seldom symmetrical, a necessary assumption for normality. Furthermore, empirical distributions tend to have heavier tails [. . .] One argument often used to justify [parametric tests] is that these tests are quite robust [. . .] [an argument] based on some old studies [about which problems exist]. (Krauth, 1988, p. 15)

Bradley concurs:

> A fantastic folk-lore sprang up among research workers: [. . .] distribution-free tests were regarded as second-class statistics, hopelessly inferior to parametric statistics [whether or not they meet their assumptions]. The efficiency of distribution-free relative to classical tests has been investigated under common non-parametric conditions, i.e. non-normal populations and/or heterogeneous variances, and the new statistics have often proven superior, sometimes infinitely so. (Bradley, 1982, p.13).

Note, too, that if you choose to use null hypotheses, those associated with distribution-free statistics are more general and thus more realistic than those of classical statistics. So if you wonder: 'When should I use distribution-free statistical methods?', the answer is: whenever possible.

Where do we go from here?

It seems likely that few of the published papers in our field meet all or even most of the requirements of scientific research (theory-oriented, conceptually clear, measurements consistent with number theory and measurement theory, preconditions for parametric statistics met or distribution-free statistics used, appropriate logic, replicative validity). As support for decisions or research, most published results and interpretations probably should be discarded. This may be too extreme but it seems to be a good starting-point.

Educational technology may be riding a wave – the wave of pseudo-science. This wave moves out of universities and through our journals. It is generated by the use of textbooks that perpetuate an unthinking, cookbook approach to scientistic rather than scientific inquiry. A radical shift in our own thinking is essential.

The challenge facing us is complex. I offer a few ideas merely as conversation starters. It strikes me that the academic preparation of educational researchers should be reformed – radically. I suggest that the research student be expected to master cybernetic principles, systems modelling, probability theory, epistemology and the philosophy of science before embarking on the study of research design and analysis. And the first stage in educational research should focus on epistemological, not statistical or even design, issues. Indeed, the first statistics course should focus on probability and non-parametric statistics with parametric statistics left for later. As for non-researchers, instead of wasting time studying cookbook-based parametric statistics, they might spend their time better studying the logic of scientific inference and the fundamentals of modelling and measurement so that they may be able to detect and reject pseudo- and quasi-pseudo-scientific research when they encounter it in our periodicals. And if they study statistics, let them begin with non-

parametric statistics which is so much easier to grasp and more generally applicable.

But this is not sufficient. journal editors and their referees need to modify their tendency to accept any quantitative article that contains a statistically significant outcome and focus more on the quality of thinking, the validity of measurement and appropriateness of qualitative and quantitative methods. Even better, they should encourage replications. Conceivably, a comparison of names of authors of pseudo-scientific articles with those on the editorial boards of our journals might reveal another problem, and journal editors may have to find new referees.

Conclusion

Much of our cherished research is pseudo-scientific with unwarranted conclusions. In considering the standard approach it is instructive to bear in mind Stafford Beer's comment:

> A paradigm is a model that exhibits a closed logic and thus resists change [. . .] To create change, you must challenge not only the models of unreality but the paradigms that underwrite them. Dangerous work. (Beer, 1988)

The paradigm informing most empirical educational research is an input-process-output model which assumes a one-way direction of causality from independent to dependent variables. This questionable model lies at the heart of t-tests, correlations, analysis of variance, multiple regression, discriminant function analysis, etc. In many papers, the mathematical prerequisites of these statistical tests are not met, thus invalidating the results.

Another fundamental flaw involves passing subjective judgements off as measurements (by attaching numbers to constructs for which none of the properties of measurable magnitude are met), then combining these, violating stringent mathematical requirements and producing meaningless results. Finally, treating linguistic variables (for example, computer-aided learning, hypertext, multimedia, learner control) as if they are precise and comparable may confuse rather than inform.

I have tried to show that our research needs to begin with conceptual issues but that our biggest problem is the way in which research is carried out. I have tried to stimulate thinking about and questioning the models and underlying paradigms that permeate our exciting field of study and practice. Whether or not you ever carry out a research programme, you are certain to encounter claims (in the mass media as well as in textbooks and journals) that are not justified or justifiable. Now you can reject them.

References

Beer, S. (1988), *Address to Convocation*, Concordia University, Montreal.

Bradley, J.V. (1982), *Distribution-Free Statistical Tests*, Englewood Cliffs, NJ: Prentice-Hall.

Entwistle, N.J. (1981), *Styles of Learning and Teaching*, Chichester: John Wiley.

Krauth, J. (1988), *Distribution-Free Statistics*, Amsterdam: Elsevier.

Lakatos, I. (1978), 'Falsification and the methodology of scientific research programmes' in Worrall, J. and Currie, G. (eds.), *The Methodology of Scientific Research Programs*, Philosophical Papers, vol. 1: Imre Lakatos, Cambridge: CUP.

Liebetrau, A.M. (1983), *Measures of Association*, London: Sage.

Lohr, L., Ross, S.M. and Morrison, G.R. (1995), 'Using a hypertext environment for teaching process writing: an evaluation study of three student groups', *Educational Technology Research and Development*, 43 (2), 33–51.

Lykken, D.T. (1968), 'Statistical significance in psychological research', *Psychological Bulletin*, 70, 151–9.

Meehl, P. (1978), 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology', *Journal of Consulting and Clinical Psychology*, 46 (4), 822.

Meredith, P. (1972), 'The origins and aims of epistemics', *Instructional Science*, 1 (1), 16.

Mitchell, P.D. (1994), 'Learning style: a critical analysis of the concept and its assessment' in Hoey, R. (ed.), *Aspects of Educational Technology XXVII*, London: Kogan Page.

Stevens, S.S. (1946), 'On the theory of scales of measurement', *Science*, 103, 677–80.